The Effect of S&P Inclusion on Stock Price

Joe Patten, Ashutosh Kumar

Abstract

The S&P 500 is an index that is comprised of 500 US companies. Every time a stock is deleted from index, the S&P 500 committee chooses eligible stocks to replace the deleted stocks. Although we know what criteria the committee looks at, it is still not clear how they choose which stock should be added. The objective of this paper is to build a classification model to predict which stocks are likely to be added each year. Using this trained classifier, we are able to build synthetic control groups which contain stocks that were likely to be chosen by the committee, but were not. We find that stock inclusion into the S&P 500 does not result in a statistically significant price increase or market capitalization increase.

Introduction

The Standard and Poor's (S&P) 500 is a stock market index that measures the performance of 500 large companies in the United States. The criteria for inclusion in the S&P 500 index is clear: stocks must be above a certain market capitalization, monthly trading volume for that stock must be above a certain value, and the stock must be from an American company that is listed on an American stock exchange. Once a stock is deleted from the S&P 500, the S&P committee gets together to pick a new stock that meets the above criteria to fill its place. Although we know the criteria that the committee looks at, it is not clear how exactly they choose which stocks are to be added. S&P states that "Company additions to and deletions from an SP equity index do not in any way reflect an opinion on the investment merits of the company" Standard and Poor's (2002). Or in other words, inclusion in the index is claimed to be an information-free event Denis et al. (2003) that should not have any effect on its price.

There has been a lot of research done to show that inclusion in the S&P 500 is associated with positive stock price increase. The earliest paper to look at if there was a bump in stock price associated with inclusion in the index was done by Shleifer (1986). Between 1966 and 1983, the authors found that over 95% of the stocks experienced an abnormally positive return right after being included. The authors concluded that from 1976 to 1983, inclusion into the index earned shareholders close to 3% capital gain in a short period of time. Chen et al. (2004) found that being included in the S&P 500 came with an increased awareness of the stock by investors, which resulted in an increased price (throughout its tenure in the index).

Denis et al. (2003) adds by showing that companies included experience a earnings per share that is significantly higher than previously expected before inclusion. The authors conclude by positing that this significant increase could be caused by a number of different reasons. One reason could be that inclusion might lead to increased scrutiny of management. Although this is probably not the case as stock price usually increases within a very short period of time (think days or weeks after a stock has been added to the index). Another reason could be that just simply being included in the index could result in the S&P "embedding" a premium on the stock. There seems to be evidence for the latter, as both Shleifer (1986) and Denis et al. (2003) have shown there to be a relatively quick increase in stock price after inclusion. Although these papers give credence to the theory that stocks receive a bump in their price after being included in the index, it is not clear how much of this increase is associated with other confounding factors. Thus, it is essential that we compare stock prices of similar stocks (which would include stocks that were included in the S&P 500, and those that barely missed the cutoff).

Recently, there has been a big push in Economics to either conduct experiments, or exploit natural experiments in order to obtain causal effects. Chang et al. (2014) use regression discontinuity in order to show that inclusion into the Russell 2000 index results in stock price increase. Regression discontinuity is a method of comparing groups that are on either side of a cutoff to be chosen for treatment. For example, one might consider the question, what effect does receiving a merit-based scholarship have on an individual? If there is a cutoff (i.e. an individual has to have a certain GPA, or must have scored a certain score on the SAT), then we could compare the outcomes of individuals who barely made the cutoff with those who barely missed the cutoff. The Russell 2000 index has a clear criteria. Based on this criteria, stocks are ranked according to a scoring method that firms are aware of. The 2000 highest ranked stocks are then included in the index. This makes it a perfect candidate for regression discontinuity. Since the criteria is clear cut, the authors were able to easily calculate which stocks were barely included in the index (which are then the treatment group), and those stocks that were barely excluded from the index (which are then the control group). Using this setup, the authors find that there is indeed a price increase when stocks are included in the Russell 2000 index.

Although there have been papers showing that there is evidence of price increase after stock inclusion in the S&P 500 (Chen et al., 2004; Shleifer, 1986; Denis et al., 2003), there are not any papers that we know of which use machine learning techniques to match similar stocks in order to infer casual effects in price change due to inclusion in the index. We hope to fill in this gap by first matching the stocks that were included in the S&P 500 with those that were likely to be included in a certain period, but were not. We will then use these matchings in order to calculate the price and market capitalization change in a stock that is a result of being included in the index. This paper will first focus on the initial objective, namely matching the stocks in each period. Then we will take the stocks that were likely to be added but were not, and compare them with stocks that were acutally added. Our hypothesis is that being added to the S&P 500's index positively impact both market capitalization and price. We first discuss different matching methods, and then propose some machine learning techniques that can be used to match similar stocks. After we have performed out group matching (and thus would have a treatment and a "synthesized" control group), we will use regression to not only see if their is an effect on price or market capitalization, but if so, how much can be attributed to being added to the index.

Matching Methods

When the S&P committee gets together, they review the criteria for S&P 500 inclusion and consider some factors before making their choice on which stock should be added. It is reasonable to believe that they consider a number of stocks before making their decision. Thus, it is not hard to believe that some stocks might have barely "missed the cutoff" in the eyes of the committee. Therefore, for every period, we want to build a model that accurately predicts both stocks that are included in the S&P 500 and those that were likely to be included, but were not. Matching techniques for causal effects has been a popular topic in the social sciences (Rubin, 2006). Propensity score matching is a method that attempts to match observations based on predictors that predict an observation (or in our case a stock) being treated (in our case being included in the S&P 500) (Rosenbaum and Rubin, 1983). The hope is that this matching will reduce the selection bias, or treatment assignment bias, that is present when trying to compare individuals that are treated with those that are not. There are other matching methods that use kernels methods and distance metrics in order to match observations (Morgan and Winship, 2015). Machine learning algorithms can also be used to match observations. Matching observations by prediction scores using a machine learning algorithm is analogous to using propensity score matching. We will also discuss how random forests can be used to extract causal effects[1].

Random Forests

Athey and Wager (2019) used random forests in order to estimate heterogenuous treatment effects in observational studies. Using data from the National Study of Learning Mindsets, which is a randomized survey done in U.S. public high schools, the authors were able to evaluate the impact of an intervention on student achievement. Their construction of random forests started from looking at the treatment effect defined by Rubin (1974). Let $Y_i^{(1)}$ and $Y_i^{(0)}$ be individual i's response to treatment and control respectively. If we know these two values, we can calculate the treatment effect at x, $\tau(x)$:

$$\tau(x) = E[Y_i^{(1)} - Y_i^{(0)} | X_i = x]$$

The problem is that when we are working with observational data, we are not able to observe both $Y_i^{(1)}$ and $Y_i^{(0)}$. Notice that we cannot estimate $\tau(x)$ without an assumption: unconfoundedness. Or in other words, treatment assignment W_i is independent of outcome $Y_i^{(j)}$. In mathematical terms, we have:

$$\{Y_i^{(0)}, Y_i^{(1)}\} \perp\!\!\!\perp W_i | X_i$$

I will first start with a regression tree, then transform it into a causal tree, and then a causal forest. Notice that we can think of trees and forests are similar nearest neighbor regressors, but we split the data and define our own neighborhoods (using some kind of entropy or gini coefficient) and take the average. Let $L(x)$ be the leaf where the test point x should be. Then our estimator is then:

$$\hat{\mu}(x) = \frac{1}{\{i \in X_i \in L(x)\}} \sum_{\{i : X_i \in L(x)\}} Y_i$$

This may look a bit scary, but this is just saying look at which leaf x should be at, and then tale the mean of the points at that leaf. Now, in order to think about this is causal terms, we will look at the function for $\tau(x)$ above, and use a similar function to define our estimator:

$$\hat{\tau}(x) = \frac{1}{\{i : W_i = 1, X_i \in L(x)\}} \sum_{\{i : W_i = 1, X_i \in L(x)\}} Y_i - \frac{1}{\{i : W_i = 0, X_i \in L(x)\}} \sum_{\{i : W_i = 0, X_i \in L(x)\}} Y_i$$

[1] We did not use random forests in the way that is explained in this paper. However, we intend use this method in future versions of this paper, thus we have chosen to include this description.

In this case, we assume that the neighborhood is small enough so it can be assumed that the (Y_i, W_i) pairs are as good as random (or in other words had come from a randomized experiment). This formula above calculates treatment effect using a tree. Trees are fairly easy to interpret, however, they can lead to overfitting. We use the definition of random forest in order to get the treatment effect using a random forest. Namely, our estimator becomes;

$$\hat{\tau}(x) = B^{-1} \sum_{b=1}^{B} \hat{\tau}_b(x)$$

where B is the number of trees in our random forest. We actually see that this treatment effect estimator has some good properties. For example, Athey and Wager (2019) showed that causal forests are consistent, and that the asymptotic variance can be accurately estimated. We can also extend random forests by thinking of them as weighted functions (similar to kernel weighting functions) (Athey and Imbens, 2019). The weighting function for a particular test point, x, is defined as:

$$\hat{\mu}(x) = \sum_{i=1}^{n} \alpha_i(x) Y_i$$

where $\sum_{i=1}^{n} \alpha_i(x) = 1$, and α's are nonnegative. Notice that for a given test point x, this counts the share of trees such that a particular observation or example is in the same leaf as x. This will weight points that are "closer"more heavily (as they are more likely to be in the same leaf as x. The top panels in figure 1 depict three different trees as well as the points in the same leaf as x. The bottom panel is a weighted function of the trees in the top panel. The more leaves that a point shares with test point x, the larger it appears (and thus has a larger weight) in the bottom panel.

Matching using Random Forests

As stated previously, trees can be seen as nearest neighbor regressors. This makes them a prime candidate for matching (Athey and Wager, 2019). By extension, a random forest is an average of matching estimators. Although random forests can be used for matching, they are not perfect. One reason they are imperfect is that they are not great at capturing linear or quadratic effects. Also, they may have problems near their "boundaries" (as there will likely have bias since the leaves of the tree cannot be centered on points near or on the boundary (Athey and Wager, 2019). One solution to this issue is to use local linear regression (i.e. regress with closer points have larger weight) (Abadie and Imbens, 2011).

Data

Monthly stock data for all US companies from 1990-2018 were taken from the Center for Research in Security Prices (CRSP) which were provided by the Wharton Research Data Services (WRDS). Current S&P 500 stocks, in addition to inclusion and deletion dates for stocks in the index were also provided by

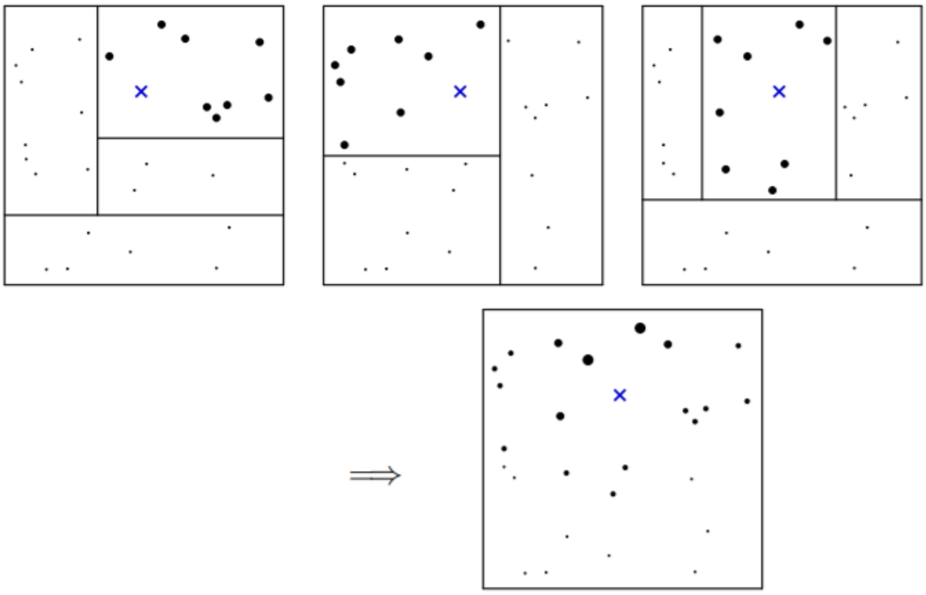

Figure 1: The top panels depict three different trees as well as the points in the same leaf as x. The bottom panel is a weighted function of the trees in the top panel. The more leaves that a point shares with test point x, the larger it appears (and thus has a larger weight) in the bottom panel (Athey and Imbens, 2019).

WRDS in the COMPUSTAT dataset. The CUSIP[2] identification variable was used in order to merge the two datasets together. Table 1 gives the summary statistics for yearly aggregated data.

	Price (in $)	Public Shares (in 1,000,000s)	Market Cap (in $1,000,000s)
count	213140	217578	213140
mean	37	65	2299
std	1551	285	13184
min	0	0	0
25%	6	5	41
50%	14	15	176
75%	28	44	804
max	306000	29049	860882

Table 1: Summary Statistics for Yearly Data

The bulk of our time was spent figuring out how we needed to construct our dataset. As our goal was to predict which stocks would be picked by the committee each period, we decided to make a dataset for each year ranging from 1990-2018. Each row of these yearly datasets were reshaped to represent a stock. Stocks that had been in the S&P 500 the previous year were taken out of that year's dataset. Also, stocks which had a market cap less then $1 billion were also taken out of that year's dataset. The features of the model were the prices, market cap, volume of shares sold, and industry type for the current and previous periods. The dependent variable was an indicator for if the stock was chosen to be added in the S&P 500 that year.

Figure 2 shows the yearly additions into the S&P 500 per year from 1990-2018. The number of

[2] CUSIP is a unique identifier for each stock. However, we ran into problems using this variable (or other similar variables) to merge data. This is discussed in the Discussion section.

additions range from 9 to 52 stocks for each year.

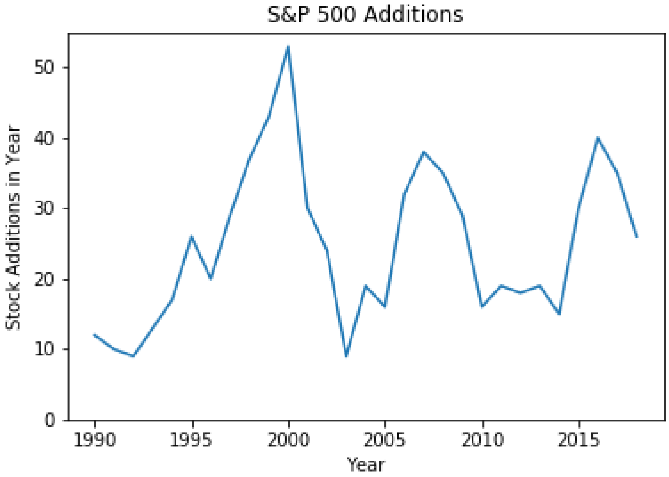

Figure 2: Stock additions to the S&P 500 per year from 1990-2018.

Results

In this paper, we will be using random forests to predict the probability of a stock being chosen by the S&P committee. The thought is that similar stocks will have similar likelihoods of being chosen, but may differ in the treatment assignment (whether they are actually included in the S&P 500). We first consider the whole dataset and see how a random forest fits the dataset. Our dependent variable in this instance is if a stock is in the S&P 500 during a certain period. We split the data into training and testing sets using an 80:20 split. After building the model, we get a training accuracy of 98.6%, and a testing accuracy of 78.04%. The confusion matrix when applying the model to the test set is shown in table 2.

	Predicted 0	Predicted 1
Actual 0	1268	478
Actual 1	868	3516

Table 2: Fitting Random Forest on all years of data. Dependent variable is if a stock is in the S&P 500. Confusion Matrix for Test Set

These are reasonable results. However, our goal is to predict additions into the index. Thus, we construct the dataset as described in the data section (where for each year we make a separate dataset). The dependent variable is now an indicator of if a stock is added to the S&P 500 that year and all other S&P 500 that were in the previous year are taken out of the dataset for that year. Table **??** shows the test accuracy, recall, and precision scores.

The accuracy is deceptively high as about 1 percent of eligible stocks were added into the S&P 500 each year (after filtering the data)[3]. Thus, we need to consider the recall and precision scores. Our goal is to have high recall scores, and lower (but still high) precision scores since we want to correctly classify stocks being added (thus recall should be high), and also find stocks that have similar probabilities as

[3] Note that since only 1 percent of the data has a label of 1, then if we predict each label to be 0, our accuracy would be 99%

	1991	1992	1993	1994	1995	1996	1997	1998	1999	2000	2001	2002	2003
accuracy	0.96	0.77	0.78	0.79	0.82	0.83	0.79	0.82	0.71	0.79	0.80	0.94	0.91
recall	1.00	1.00	0.60	1.00	0.80	1.00	0.89	0.64	0.43	0.57	0.44	0.20	0.29
precision	0.14	0.05	0.08	0.11	0.08	0.10	0.11	0.09	0.06	0.06	0.07	0.05	0.05

	2004	2005	2006	2007	2008	2009	2010	2011	2012	2013	2014	2015	2016
accuracy	0.87	0.89	0.83	0.77	0.78	0.88	0.90	0.91	0.91	0.91	0.87	0.86	0.90
recall	0.50	0.42	0.56	0.33	0.75	0.14	0.33	0.50	0.33	0.50	0.67	0.57	0.29
precision	0.03	0.08	0.05	0.04	0.03	0.02	0.04	0.04	0.03	0.03	0.09	0.04	0.06

Table 3: Test Accuracy, precision, and recall scores for the random forests models fit on each year's training data. Dependent variable is if a stock is added to the S&P 500 that year.

being predicted as being added that year to the index.

Recall scores are fairly high at least for the first few years, but tend to get lower the closer we get to 2016. Precision scores are lower than we'd like, in other words we have too many false positives. We have tweaked the model by performing hyperparameter tuning, and still get similar results. Also, resampling methods were used so that we would have a comparable portion of positive labels in our training set as negative labels. However, this did not affect accuracy, recall or precision scores. As we want to get stocks that were likely to be added, we have fit a random forest on each dataset (not just the testing portion) in order to do the matching.

We will use the following model to evaluate how price and market share are affected by S&P 500 inclusion for stock i:

$$Y_i = \beta_0 + \beta_1 \mathbb{I}(add_i) + \gamma_t \tag{1}$$

where γ_t year fixed effect (for year t), $Y_i \in \{\Delta \log(price_i),\ \Delta \log(marketcap_i)\}$, and $\mathbb{I}(add_i)$ is the indicator for whether stock i was added or not that period. β_1 is our parameter of interest, as it will tell us the effect of inclusion into S&P 500 on stock price or market capitalization[4]. To make this problem a bit easier, I have made each time period be in terms of years. Thus, we are looking at the effect of being added in the S&P 500 this year on changes in price and market capitalization from this year to next year. Table 4 shows the result for equation 4.

	$\Delta \log(price_i)$	$\Delta \log(marketcap_i)$
Added to S&P 500	.0057	.0258
	(.1025)	(.217)

Table 4: Although both coefficients are positive (which is what was to be expected), neither are significant at the .05 level.

Each column in the table above represents one regression. You will notice that although both coefficients are positive. However, they are both not statistically significant. Thus we cannot reject the null hypothesis that addition into the S&P 500 has no effect on price or market capitalization.

[4] Obviously, the assumption being that the stocks we predicted to be added in the S&P 500 for that period are similar to the stocks actually being added in that period

Discussion

We want to address our results as well as some major concerns or drawbacks. First, we ran into a problem of being able to uniquely match stocks between the monthly stock data from CRSP, and the S&P 500 current members dataset from COMPUSTAT. This resulted in 2 problems. First, not being able to identify all of S&P 500 current members means that we were not able to take out all of the current S&P 500 members from each yearly dataset. This creates a problem when trying to predict which stocks will be picked by the committee. If a stock is in the S&P 500, it is more than likely the case that that stock's features have much higher values (for example market capitalization) than stocks outside of the index. Our classifier has to deal with this, which may unintentionally leaded to bias. Another problem is that we might not be identifying all the stocks that are to be added in a certain period. This would result in similar consequences, namely that our classifier will be biased since some of our labels could be wrong.

Since we divided data by year, it was natural to look at the change in log price and the change in log market capitalization over that year. We found that there was not a statistically significant (at the .05 level) increase in either of these two variables. This is consistent with some of the literature (Shleifer, 1986; Denis et al., 2003), which found that inclusion in the S&P 500 only resulted in a temporary increase in stock price or market capitalization. Thus, we need to look at a more "finer" time scale. In the future, we hope to obtain daily and weekly stock data for companies added to the S&P 500 and those that were likely to be added. Then we want to compare stock price and market capitalization in the days and weeks before a company is added, and the days and weeks after is has been added. We suspect that there would be a significant bump in price and market capitalization when looking at this fine level of detail.

As stated previously, the bulk of our time was spent figuring out how to structure our data, and then actually executing out plans. We started out with a very large dataset containing almost 30 years of monthly stock data for tens of thousands of companies. We would have likes to add more features into both the matching and causal estimation steps. Some of the features might include financial statements from each company, characteristics of a company, and other information along those lines. We searched for data, and also contemplated scraping data from multiple sources, however, it became too large of a task for this project. In the future we hope to add these or similar features in order to increase accuracy in the matching stage, and thus give better estimates in the causal stage.

In the future, we also want to compare various matching techniques to see how our results change based on the matching. In this paper, we chose build build a classifier that would predict inclusion into the S&P 500 in order to make a "synthetic" control group for each year. This approach made sense to us as we were able to predict stocks that were likely to get (and thus were similar to stocks that got in), which made a great synthetic control group. However, we would like to experiment with other matching techniques such as using random forests to essentially get a nearest neighbor approach as modelled by Athey and Wager (2019) or by using clustering to group similar stocks (Athey and Mobius, 2012).

Conclusion

We have built classifiers using random forests in order to predict whether a stock is likely to be added into the S&P 500. This approach is similar to what is done when using propensity score matching which is usually used in the social sciences. Our hope is that we were able to match stocks that were not added to the S&P 500, but were similar to those stocks that were added into the index (and basically only differ in that respect). Using this matching, we were able to construct a control group, which contained stocks that were likely to be added during a certain period but were not, and a treatment group, which contained stocks that were actually added during a certain period. Using these two groups, we have constructed a causal framework in order to evaluate the effect of addition into the S&P 500 on a stock's price and market capitalization. We did not find a statistically significant effect of this inclusion on price change or market capitalization change when focusing on yearly change. This is consistent with what was found by Shleifer (1986); Denis et al. (2003). We are not entirely sure that this is actually the case, as some of our data is mislabelled (as discussed in the Discussion section). In the future, we would want to correct this error, and redo our analysis. We would also like to look at a smaller time frame for the effect of inclusion on price and market capitalization, possibly days, weeks or months.

References

Abadie, A. and Imbens, G. W. (2011). Bias-corrected matching estimators for average treatment effects. *Journal of Business & Economic Statistics*, 29(1):1–11.

Athey, S. and Imbens, G. W. (2019). Machine learning methods that economists should know about. *Annual Review of Economics*, 11.

Athey, S. and Mobius, M. (2012). The impact of news aggregators on internet news consumption: The case of localization. In *Workshop on the Economics of Web Search and Social Network. Sixth ACM International Conference on Web Search and Data Mining 2013*, page 2.

Athey, S. and Wager, S. (2019). Estimating treatment effects with causal forests: An application. *arXiv preprint arXiv:1902.07409*.

Chang, Y.-C., Hong, H., and Liskovich, I. (2014). Regression discontinuity and the price effects of stock market indexing. *The Review of Financial Studies*, 28(1):212–246.

Chen, H., Noronha, G., and Singal, V. (2004). The price response to s&p 500 index additions and deletions: Evidence of asymmetry and a new explanation. *The Journal of Finance*, 59(4):1901–1930.

Denis, D. K., McConnell, J. J., Ovtchinnikov, A. V., and Yu, Y. (2003). S&p 500 index additions and earnings expectations. *The Journal of Finance*, 58(5):1821–1840.

Morgan, S. L. and Winship, C. (2015). *Counterfactuals and causal inference*. Cambridge University Press.

Rosenbaum, P. R. and Rubin, D. B. (1983). The central role of the propensity score in observational studies for causal effects. *Biometrika*, 70(1):41–55.

Rubin, D. B. (1974). Estimating causal effects of treatments in randomized and nonrandomized studies. *Journal of educational Psychology*, 66(5):688.

Rubin, D. B. (2006). *Matched sampling for causal effects*. Cambridge University Press.

Shleifer, A. (1986). Do demand curves for stocks slope down? *The Journal of Finance*, 41(3):579–590.

Standard and Poor's (2002). Focusing the s&p 500 on us large cap stocks and the removal of non-us companies in the s&p 500.

Code Appendix

Code for this project can be found at the following github repo:
`https://github.com/joepatten/Data-Science-Final-Project/blob/master/README.md`